UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN

Digitized by the Internet Archive
in 2013

http://archive.org/details/useofriccatiequa674triv

Illin
no.674 UIUCDCS-R-74-674
Cop 2

The Use of Riccati Equation in Digital Computer Arithmetic

by

Kishor S. Trivedi

August 1974

THE LIBRARY OF THE
OCT 4 1974
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

N · URBANA, ILLINOIS

This volume is bound without no.676,vol.I,A-L.

which is/are unavailable.

Illn
no.674
Cop 2

UIUCDCS-R-74-674

The Use of Riccati Equation in Digital Computer Arithmetic

by

Kishor S. Trivedi

August 1974

THE LIBRARY OF THE
OCT 4 1974
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

UIUCDCS-R-74-674

The Use of Riccati Equation in Digital Computer Arithmetic

by

Kishor S. Trivedi

August 1974

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois

This work was supported in part by the National Science Foundation under Grant No. US NSF GJ 38204.

510.84
I16N
no.674-678
Cop 2

1. INTRODUCTION

It was demonstrated by DeLugish [1] that number representations other than the traditional positional notation are useful. This led to a search for new number representations.

There seem to be several basic requirements for a proposed representation of numbers to be useful for implementation in hardware [2]. These are the following:

Requirement 1: Conversion to the conventional positional notation should be simple.

Requirement 2: The set of numbers representable should be a complete set [3]. In other words, the range of numbers representable should be an interval of the real numbers.

Requirement 3: It should be possible to define a broad class of algorithms that are easily soluble for the representation of numbers used. Hardware-compatibility among the algorithms is also desirable.

Requirement 4: For algorithms which require trial and error procedure, it should be possible to devise techniques such that the coefficient-selection is practical (cf., quotient-digit selection in division).

For continued fractions, a simple conversion procedure to positional notation is well known [2,4]. By an appropriate choice of the digit sets for the partial numerators and partial denominators, the requirement of completeness is satisfied [2,5]. The requirement of the identification of a broad class of algorithms seems to be a difficult problem. A limited class of quadratics can be solved using continued fractions [2]. We will

discuss a very powerful method by which continued fraction algorithms for many elementary functions and for the integration of many functions can be obtained. The requirement of an appropriate selection procedure is another difficult problem. For the solution of quadratics, a selection procedure has been obtained [2,5].

The use of Riccati equation for continued fraction arithmetic will be discussed in this report. We will first show the large number of functions that fit into this analysis. Later we will discuss the derivation of suitable selection procedures. The use of Riccati equation for this purpose was first proposed in [6].

2. RICCATI EQUATION

2.0 Introduction

Riccati equation can be written as:

$$y' + a(x)y^2 + b(x)y + c(x) = 0. \tag{2.1}$$

Let L be the set of all Riccati equations of this form. Wynn has shown that the set L is closed under the bilinear transformation $y = p/(q+z)$ where p,q are constants [7]. Starting with $\ell_0 \in L$, by a repeated application of the bilinear transformation, we can obtain a continued fraction expansion for the solution to the initial Riccati equation ℓ_0.

Let ℓ_0 be given by: $y_0' = a_0 y_0^2 + b_0 y_0 + c_0$, and let $y_0 = p_1/(q_1+y_1)$. Let this transformation be called T_1: $L \to L$. $T_1(\ell_0) = \ell_1$ is given by $y_1' + a_1 y_1^2 + b_1 y_1 + c_1 = 0$.

The recursion relations for the coefficients a_1, b_1, c_1 in terms of a_0, b_0, c_0 are,

$$\left.\begin{aligned} a_1 &= c_0/p_1, \\ b_1 &= b_0 + 2\, c_0\, q_1/p_1, \\ c_1 &= a_0\, p_1 + b_0\, q_1 + c_0\, q_1^2/p_1 \end{aligned}\right\} \tag{2.2}$$

Note here that, we have changed the form of ℓ_0 to avoid negative signs in recursions (2.2). In general, let $\ell_{2m} = (y_{2m}' = a_{2m}\, y_{2m}^2 + b_{2m}\, y_{2m} + c_{2m})$ and let $\ell_{2m+1} = (y_{2m+1}' + a_{2m+1}\, y_{2m+1}^2 + b_{2m+1}\, y_{2m+1} + c_{2m+1} = 0)$. Assume that,

$\ell_n = T_n\, T_{n-1} \cdots T_1(\ell_0)$ has been obtained. Then the coefficients of $\ell_{n+1} = T_{n+1}(\ell_n)$ are given by the following recursions:

$$\left.\begin{aligned} a_{n+1} &= c_n/p_{n+1} \\ b_{n+1} &= b_n + 2\, c_n\, q_{n+1}/p_{n+1} \\ c_{n+1} &= a_n\, p_{n+1} + b_n\, q_{n+1} + c_n\, q_{n+1}^2/p_{n+1}. \end{aligned}\right\} \tag{2.3}$$

As a result of these transformations, we have expanded y_0 to n+1 terms as follows:

$$y_0 = \frac{p_1}{q_1} + \frac{p_2}{q_2} + \cdots\cdots\cdots + \frac{p_{n+1}}{q_{n+1}+y_{n+1}}\,. \tag{2.4}$$

Let P_{n+1}/Q_{n+1} denote the finite continued fraction obtained by setting $y_{n+1} = 0$ in equation (2.4). If we assume that $|y_n| < M$ where M is a fixed constant then clearly, the fraction P_n/Q_n converges to y_0. By setting, $P_0 = 0$, $Q_0 = 1$, $P_1 = p_1$ and $Q_1 = q_1$, the recursions for P_{n+1} and Q_{n+1} are [4]:

$$\left.\begin{aligned} P_{n+1} &= q_{n+1}\, P_n + p_{n+1}\, P_{n-1} \\ Q_{n+1} &= q_{n+1}\, Q_n + p_{n+1}\, Q_{n-1} \end{aligned}\right\} \tag{2.5}$$

Thus if we have a method to correctly choose p_n, q_n for every n then, we have an algorithm to solve the Riccati equation.

2.1 Power of the Method

We will now discuss the number of functions that can be obtained by the method of Riccati equation.

2.1.1 Constant Coefficients

Let us consider a subset L_0 of L such that $L_0 = \{y' + ay^2 + by + c = 0 \mid a, b, c \in R\}$ i.e., the set of all Riccati equations with constant coefficients. Consider $\ell_0 \in L_0$ given by, $y_0' = a_0 y_0^2 + b_0 y_0 + c_0$. Depending on the sign of $\Delta = b_0^2 - 4 a_0 c_0$, the solution $y_0(x)$ of ℓ_0 can be written as,

$$y_0(x) = \frac{\sqrt{-\Delta}}{2a_0}\left(\tan\left(\frac{\sqrt{-\Delta}}{2} x + A_0\right) - \frac{b_0}{\sqrt{-\Delta}}\right)$$

if $\Delta < 0$ and $a_0 \neq 0$;

$$y_0(x) = -\frac{1}{a_0 x} - \frac{b_0}{2a_0} + A_0 \qquad \text{if } \Delta = 0,\ a_0 \neq 0;$$

$$y_0(x) = \frac{\sqrt{\Delta}}{-2a_0}\left(\tan h\left(\frac{\sqrt{\Delta}}{-2} x + A_0\right) - \frac{b_0}{\sqrt{\Delta}}\right)$$

if $\Delta > 0$, $a_0 \neq 0$;

$$y_0(x) = A_0 e^{b_0 x} + c_0 x \qquad \text{if } a_0 = 0.$$

Depending on the values of the coefficients a_0, b_0, c_0 and the initial condition $t_0 = y_0(0)$, many different functions may be evaluated as shown in the following table.

a_0	b_0	c_0	Δ	t_0	$y_0(x)$
1	0	1	-4	0	tan x
-1	0	-1	-4	∞	cot x
-1	0	0	0	∞	1/x
-1	0	1	4	∞	cot h x
-1	0	1	4	0	tan h x
0	± 1	0	>0	1	$e^{\pm x}$

Table 2.1

2.1.2 Variable Coefficients

Consider a subset L_1 of L so that,

$$L_1 = \{y' = a(x)\, y^2 + b(x)\, y + c(x) \mid a(x) = k(x)\, \bar{a},$$

$$b(x) = k(x)\, \bar{b},\ c(x) = k(x)\, \bar{c},\ \text{and } \bar{a},\ \bar{b},\ \bar{c}$$

$$\text{are constants}\}.$$

Recursions for $\bar{a}_{n+1}$, $\bar{b}_{n+1}$ and $\bar{c}_{n+1}$ can be derived from the recursions (2.3) and are as follows:

$$\left.\begin{aligned} \bar{a}_{i+1} &= \bar{c}_i / p_{i+1} \\ \bar{b}_{i+1} &= \bar{b}_i + 2\bar{c}_i\, q_{i+1}/p_{i+1} \\ \bar{c}_{i+1} &= \bar{a}_i\, p_{i+1} + \bar{b}_i\, q_{i+1} + \bar{c}_i\, q_{i+1}^2/p_{i+1}. \end{aligned}\right\} \qquad (2.6)$$

Depending on the sign of $\bar{\Delta}_0 = \bar{b}_0^2 - 4\bar{a}_0\, \bar{c}_0$, the solution to ℓ_0 is given by:

$$y_0(x) = \frac{\sqrt{-\bar{\Delta}_0}}{2\bar{a}_0}\left(\tan\left(\frac{\sqrt{-\bar{\Delta}_0}}{2}\int k(x)\,dx + A_0\right) - \frac{\bar{b}_0}{\sqrt{-\bar{\Delta}_0}}\right)$$

$$\text{if } \bar{\Delta}_0 < 0,\ \bar{a}_0 \neq 0;$$

$$y_0(x) = -\frac{1}{\bar{a}_0 \int k(x)\,dx} - \frac{\bar{b}_0}{2\bar{a}_0} + A_0$$

$$\text{if } \bar{\Delta}_0 = 0,\ \bar{a}_0 \neq 0;$$

$$y_0(x) = -\frac{\sqrt{\bar{\Delta}_0}}{2\bar{a}_0}\left(\tan h\left(\frac{\sqrt{\bar{\Delta}_0}}{-2}\int k(x)\,dx + A_0\right) - \frac{\bar{b}_0}{\sqrt{\bar{\Delta}_0}}\right)$$

$$\text{if } \bar{\Delta}_0 > 0,\ \bar{a}_0 \neq 0;$$

$$y_0(x) = A_0\, e^{\bar{b}_0 \int k(x)\,dx} - \frac{\bar{c}_0}{\bar{b}_0} \quad \text{if } a_0 = 0,\ \bar{b}_0 \neq 0;$$

and

$$y_0(x) = \bar{c}_0 \int k(x)\,dx + A_0 \quad \text{if } \bar{a}_0 = \bar{b}_0 = 0.$$

Clearly, a large class of functions can be evaluated with this method.

2.1.2.1 The Case With $\bar{\Delta}_0 = 0$

In this section, we will concentrate on a subset L_{10} of L_1 such that, $L_{10} = \{\ell \in L_1 | \bar{\Delta}_0 = 0\}$. Any $\ell \in L_{10}$ can be rewritten as: $y' = k(x)(a^*y+b^*)^2$ where, $a^* = \sqrt{\bar{a}}$, $b^* = a^*(\frac{\bar{b}}{2\bar{a}})$. With this modification, we have reduced the number of coefficients from three to two. The recursions on a^*_n, b^*_n can now be written as follows:

$$\left.\begin{aligned} a^*_{n+1} &= b^*_n/\sqrt{p_{n+1}}, \\ b^*_{n+1} &= (a^*_n\, p_{n+1} + b^*_n\, q_{n+1})/\sqrt{p_{n+1}} \end{aligned}\right\} \qquad (2.7)$$

The solution $\ell_0 \in L_{10}$ is given by,

$$y_0(x) = \frac{1}{(a^*_0)^2(A_0 - \int k(x)\,dx)} - \frac{b^*_0}{a^*_0} \quad \text{if } a^*_0 \neq 0,$$

$$y_0(x) = (b^*_0)^2 \int k(x)\,dx + A_0 \quad \text{if } a^*_0 = 0.$$

Note that, we can integrate the given function k(x) by this method by setting $a^*_0 = 0$ and $b^*_0 = 1$.

2.2 Implementation Considerations

Let us assume that simple selection procedures are available for all the functions to be evaluated as detailed in section 2.1. We now give steps of an algorithm T which will evaluate these functions.

Algorithm T:

Step 1: [Initialize]

Set $P_0 \leftarrow 0$, $Q_0 \leftarrow 1$, $P_{-1} \leftarrow 1$, $Q_{-1} \leftarrow 0$;

Set initial values of coefficients according

to the function to be evaluated;

Set $i \leftarrow 0$;

Step 2: [Select]

$(p_{i+1}, q_{i+1}) \leftarrow$ Select(x, coefficients, function);

Step 3: [Recursions]

$P_{i+1} \leftarrow q_{i+1} P_i + p_{i+1} P_{i-1}$;

$Q_{i+1} \leftarrow q_{i+1} Q_i + p_{i+1} Q_{i-1}$;

Recurse using equations (2.3), (2.6) or

(2.7) whichever is applicable.

Step 4: [Test]

After 'sufficient' number of iterations

GO TO Step 5; otherwise set $i \leftarrow i+1$,

and GO TO Step 2;

Step 5: [evaluate]

$y_0(x) = f(x) = P_{i+1}/Q_{i+1}$;

END T;

In any such iterative algorithm, the number of iterations required and the execution time required per iteration are two important considerations.

In each iteration, steps 2, 3 and 4 are executed. Clearly, step 2 and 3 require more attention. We can assume that if the procedure Select is known, it can be implemented in a combinational network and therefore, will require very little time. In step 3, we see that all the assignments are independent of each other and therefore, can be executed in parallel. Thus, given sufficient hardware, step 3 can be speeded up considerably. Each individual recursion requires additions (subtraction), multiplications and sometimes division also. Since multiplication and division are relatively slower operations, we would like to avoid them if possible. If we restrict the coefficients p_i and q_i to be integral powers of two these multiplications and divisions will be reduced to shifts, which is relatively a faster operation. If we use the recursions (2.7), then we further require that $p_i = 1$ for $\forall$ i. For the sake of simplicity, we will assume that $p_i = 1$ in every case and $q_i \in S_q$ where, $S_q = \{2^j | j \text{ is an integer}\}$. Since the number of shifts available is finite we further require that, $S_q = \{2^j | \underline{J}_q \leq j \leq \overline{J}_q\}$ where $\overline{J}_q$ and $\underline{J}_q$ are fixed integers.

The number of iterations to be carried out can be decided on the basis of allowable error in the result.

2.3 Initial Condition

Associated with the solution of any differential equation, there are one or more arbitrary constants which are evaluated using the boundary conditions imposed. Depending on the Function f(x) to be evaluated, we choose a particular $\ell_0 \in L$ (and the corresponding coefficient values) and the associated initial condition $y_0(0) = t_0$ so that $y_0(x) = f(x)$. Clearly, the initial condition on ℓ_i $(i \geq 1)$ is dependent on t_0. In particular,

$$y_{n-1} = p_n/(q_n+y_n) \text{ which implies,}$$
$$t_{n-1} = p_n/(q_n+t_n) \text{ which implies,}$$
$$t_n = p_n/t_{n-1} - q_n \qquad (2.8)$$

As we will see in Chapter 3, t_n is needed as an argument in a selection procedure for p_{n+1} and q_{n+1}. Therefore, we need to evaluate t_n in every iteration. This, however, implies that a division be carried out. We can avoid the division by the following technique.

Let $t_n = d_n/e_n$ then, from equation (2.8),

$$\frac{d_n}{e_n} = - q_n + \frac{p_n\, e_{n-1}}{d_{n-1}}$$

From which,

$$\left.\begin{aligned} d_n &= p_n\, e_{n-1} - q_n\, d_{n-1} \\ e_n &= d_{n-1} \end{aligned}\right\} \qquad (2.9)$$

and $\qquad d_0 = t_0 \qquad$ and $\qquad e_0 = 1.$

If the selection procedure can choose with the help of d_n and e_n (does not explicitly require t_n) then we have solved our problem. Now in step 3 of algorithm T, we have to carry out recursions (2.9) as well.

2.4 Convergence

If the digit sets S_p and S_q are finite and positive then it can easily be shown that $m \leq y_n \leq M \;\forall\; n$, i.e., y_n is bounded [5]. The convergence P_n/Q_n to $y_0(x)$ can then be guaranteed if a correct selection procedure is available [5]. Therefore, we will discuss the question of selection procedure in the next chapter.

3. ATTEMPTS AT FINDING SELECTION PROCEDURES

3.0 Introduction

It is quite clear that the selection procedure will be different for different functions. We will say that a selection procedure is consistent if at every step in the evaluation of a given function, the same selection procedure can be used.

The definition of consistency implies that the function represented by $y_i(x)$ is independent of i. From the solutions of Riccati equation (given in Chapter 2), it is easy to see that this will be true if $\Delta_i = \Delta$ for $\forall$ i. This can be shown as follows [8]: Let $\ell_{i-1} \in L$ and $T_i(\ell_{i-1}) = \ell_i$ then from recursions (2.3), we have,

$$\Delta_i = b_i^2 - 4 a_i c_i$$

$$= (b_{i-1} + c_{i-1} q_i/p_i)^2$$

$$- 4 \frac{c_{i-1}}{p_i} (a_{i-1} p_i + b_{i-1} q_i + c_{i-1} q_i^2/p_i)$$

$$= b_{i-1}^2 + 4 c_{i-1}^2 q_i^2/p_i^2 + 4 b_{i-1} c_{i-1} q_i/p_i$$

$$- 4 a_{i-1} c_{i-1} - 4 b_{i-1} c_{i-1} q_i/p_i - 4 c_{i-1}^2 q_i^2/p_i^2$$

$$= b_{i-1}^2 - 4 a_{i-1} c_{i-1}$$

$$= \Delta_{i-1}.$$

Thus we have shown that consistency requirement is satisfied.

We now fix the digit sets for the partial numerators and the partial denominators. We will assume $p_i = 1$ for all i. We will assume $S_q = \{1, 1/2, 1/4\}$. Note that this is a redundant digit set. Such a redundancy is required for the selection procedure to be possible [5].

The set of all infinite continued fractions representable with this digit set is the interval $F = [0.39, 1.56]$. Given $y_i \in F$ and $y_i = \dfrac{1}{q_{i+1}+y_{i+1}}$, the constraint for choosing q_{i+1} is that $q_{i+1} \in S_q$ and $y_{i+1} \in F$. Therefore, if $y_i \in F_1 = [0.39, 0.72]$ then we can choose $q_{i+1} = 1$, if $y_i \in F_2 = [0.485, 1.123]$ then we can choose $q_{i+1} = 1/2$, and if $y_i \in F_3 = [0.553, 1.56]$ then we can choose $q_{i+1} = 1/4$. Quite clearly, the intervals F_1, F_2 and F_3 are not pairwise mutually disjoint, which shows the redundancy of the digit set S_q. In the intersection $F_{12} = F_1 \cap F_2$, q_{i+1} can be chosen as either 1 or 1/2. Similarly, in the interval $F_{23} = F_2 \cap F_3$, q_{i+1} can be chosen as either 1/2 or 1/4. Thus, if we know y_0 then we have a consistent method of expansion of y_0 into an infinite continued fraction. Since y_0 is not known in advance, this simple method will not work.

3.1 The Criteria for Selection

For further discussion, we will concentrate on the subset L_{10} of L. Consider $\ell \in L_{10}$ given by: $y_n' = j\, k(x)\, (a_n y_n + b_n)^2$ where $j = 1$ if n is even and $j = -1$ if n is odd. The solution to this equation is given by:

$$A_n = j\, g(x) + \frac{1}{a_n(a_n y_n + b_n)} \qquad (3.1)$$

Where, A_n is an arbitrary constant of integration and $g(x) = \int k(x)\, dx$. A_n can be evaluated by using the initial condition $y_n(0) = t_n = \dfrac{d_n}{e_n}$. Therefore, $A_n = j\, g(0) + \dfrac{e_n}{d_n} \dfrac{1}{(a_n d_n + b_n e_n)}$. Substituting this value of A_n in equation (3.1),

we have,

$$y_n(x) = \frac{d_n + j(g(x)-g(0))\, b_n(a_n d_n + b_n e_n)}{e_n - j(g(x)-g(0))\, a_n(a_n d_n + b_n e_n)}$$

$$\triangleq S_n(k(x), a_n, b_n, d_n, e_n, j). \qquad (3.3)$$

Quite clearly, if the function S_n can be evaluated accurately for every value of n then the selection procedure is trivial. But the computation of the function S_n is at least as complex as that of $y_0(x)$. Therefore, this method is not useful. However, the redundant number representation may allow us to use an approximation $\hat{S}_n$ of S_n in the selection procedure. If the computation of $\hat{S}_n$ is simple enough then such a selection procedure will be practical.

Let us denote $\bar{z}_{12} = \sup.(F_{12})$, $\underline{z}_{12} = \text{Inf.}(F_{12})$, $\bar{z}_{23} = \text{Sup.}(F_{23})$ and $\underline{z}_{23} = \text{Inf.}(F_{23})$. Further let z_1 and z_2 be suitable constants such that $z_1 \in F_{12}$, $z_2 = F_{23}$. We use the following selection procedure, Select 1.

Select 1:

If $z_1 \leq \hat{S}_n \leq z_2$ then $q_{i+1} = 1/2$;

else

if $\hat{S}_n > z_2$ then $q_{i+1} = 1/4$;

else

if $\hat{S}_n < z_1$ then $q_{i+1} = 1$;

end Select 1;

Clearly, Select 1 will be correct if the following condition is satisfied for all n. $|\hat{S}_n - S_n| \leq \text{Min.}\{\bar{z}_{12} - z_1, \bar{z}_{23} - z_2, z_1 - \underline{z}_{12}, z_2 - \underline{z}_{23}\}$.

.... (3.4)

Therefore, we should try to find a sufficiently simple approximation $\hat{S}_n$ for S_n which satisfies the condition (3.4).

A linear approximation to S_n is perhaps the simplest. Further, if the coefficients in the linear approximation are powers of two then computation of $\hat{S}_n$ will involve addition and shift operations only.

It may be noted that with $S_q = \{1, 1/2, 1/4\}$, $F_{12} = [0.485, 0.72]$ and $F_{23} = [0.553, 1.123]$. If we let $z_1 = 0.5$ and $z_2 = 0.75$, then even if $\hat{S}_n$ is correct only to one digit, q_{n+1} can be chosen correctly.

Before proceeding further, we choose a particular function to be evaluated. We let $y_0(x) = \mathrm{Sin}^{-1}(x)$, $k(x) = 1/\sqrt{1-x^2}$, $a_0 = 0$, $b_0 = 1$, $d_0 = 0$, $e_0 = 1$.

3.1.1 More Details

For $y_0(x) = \mathrm{Sin}^{-1}(x)$,

$$S_n = \frac{d_n + j\,\mathrm{Sin}^{-1}(x)\, b_n(a_n d_n + b_n e_n)}{e_n - j\,\mathrm{Sin}^{-1}(x)\, a_n(a_n d_n + b_n e_n)} \tag{3.5}$$

Therefore, $S_n \in F_{12}$ iff,

$$jx \leq \mathrm{Sin}\left(\frac{(0.72\, e_n - d_n)}{0.72\, a_n^2\, d_n + (0.72\, e_n + d_n)(a_n b_n) + b_n^2 e_n}\right)$$

and

$$jx \geq \mathrm{Sin}\left(\frac{(0.485\, e_n - d_n)}{0.485\, a_n^2\, d_n + (0.485\, e_n + d_n)\, a_n\, b_n + b_n^2 e_n}\right) \tag{3.6}$$

Let

$$M_n(s) = \mathrm{Sin}\left(\frac{s\, e_n - d_n}{s\, a_n^2\, d_n + (s\, e_n + d_n)\, a_n\, b_n + b_n^2 e_n}\right)$$

The selection overlap regions can be given by:

$$jx \in [M_n(0.485),\ M_n(0,72)] \quad \text{then} \quad q_{n+1} = 1/2 \text{ or } 1,$$

and

$$jx \in [M_n(0.553),\ M_n(1.123)] \quad \text{then} \quad q_{n+1} = 1/4 \text{ or } 1/2.$$

Since the computation involved in $M_n(s)$ is intolerable, we must find a simple approximation $\hat{M}_n(s)$ to $M_n(s)$. If $\hat{M}_n(s)$ satisfies the condition (3.4) then the selection procedure will be correct. There are several ways to obtain the approximation $\hat{M}_n(s)$.

The first attempt consisted of trying to obtain a continuous linear least square fit for $jx = M_n(s)$ which is linear in a_n, b_n, d_n, e_n. But this procedure involved the computation of a number of 4-tuple integrals of functions which are inseparable in the variables of integration. Therefore, this approach was abaondoned.

Second approach consisted of obtaining a tangential hyperplane to the surface $jx = M_n(s)$ (for a fixed value of s) at some point in the five dimensional space R^5. Unfotunately, after some computation, it was seen that for displacements in the 5-space, the tagential hyperplanes rotate by large angles.

The third approach consisted in obtaining a discrete least square fit of $jx = M_n(s)$. This approach also could not produce an approximation which will satisfy the error condition (3.4).

The growth of coefficients a_n, b_n, d_n, e_n seems to be the reason of failure.

4. CONCLUSION AND FUTURE RESEARCH

The method of Riccati equation is a very powerful method. Countless number of functions can be evaluated in a uniform manner using this method. Obtaining a suitable selection procedure seems to be a very difficult problem.

Further investigation for obtaining selection procedures is needed. Growth pattern of the coefficients of the Riccati equation need to be studied. Methods of scaling so as to control the growth of coefficients need to be investigated.

LIST OF REFERENCES

[1] DeLugish, B. G., "A Class of Algorithms for Automatic Evaluation of Certain Elementary Functions in a Binary Computer," Ph.D. Dissertation, University of Illinois, Urbana, June 1970; also Department of Computer Science Report 399.

[2] Robertson, J. E. and K. S. Trivedi, "The Status of Investigations into Computer Hardware Design Based on the Use of Continued Fractions," IEEE Transactions of Computers, Vol. C-22, No. 6, June 1973, pp. 555-560.

[3] Goldberg, R. R., "Methods of Real Analysis," Blaisdell Publishing Company, New York, 1965.

[4] Wall, H., "Analytic Theory of Continued Fractions," Van Nostrand, Princeton, New Jersey, 1950.

[5] Trivedi, K. S., "An Algorithm for the Solution of a Quadratic Equation using Continued Fractions," M.S. Thesis, University of Illinois, Urbana, June 1972; also Department of Computer Science Report 525.

[6] Bracha, A., "Application of Certain Functions on a Digital Computer," University of Illinois, Urbana, Technical Report, UIUCDCS-R-72-510.

[7] Wynn, P., "On Some Recent Developments in the Theory and Application of Continued Fractions," Journal SIAM on Numerical Analysis, Vol. 1, pp. 177-197, 1964.

[8] Robertson, J. E., private communication.

BIBLIOGRAPHIC DATA SHEET	1. Report No. UIUCDCS-R-74-674	2.	3. Recipient's Accession No.

4. Title and Subtitle

The Use of Riccati Equation in Digital Computer Arithmetic

5. Report Date: August 1974

6.

7. Author(s)

Kishor S. Trivedi

8. Performing Organization Rept. No.

9. Performing Organization Name and Address

Department of Computer Science
University of Illinois
Urbana, Illinois

10. Project/Task/Work Unit No.

11. Contract/Grant No.

GJ 38204

12. Sponsoring Organization Name and Address

National Science Foundation
Washington, DC

13. Type of Report & Period Covered

14.

15. Supplementary Notes

16. Abstracts

This is a report of ongoing research in continued fraction arithmetic. It is shown that a large number of functions can be evaluated using the Riccati differential equation. However, the problem of properly selecting the partial numerators and the partial denominators remains to be solved. Several attempts to obtain such selection procedures are described. Possible directions for further research are indicated.

17. Key Words and Document Analysis. 17a. Descriptors

Combinational Network
Computer Arithmetic

Continued Fractions
Function Evaluation
Representation of Numbers

Riccati Equation

Selection Procedure

17b. Identifiers/Open-Ended Terms

17c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages
	20. Security Class (This Page UNCLASSIFIED	22. Price

FEB 20 1991

BOUND BY HM

UNIVERSITY OF ILLINOIS-URBANA
510.84 IL6R no. C002 no.674-678(1974
Use of Riccati equation in digital compu

3 0112 088401465